VeriCrux Protocol v2.8 – Sovereign Truth Discovery Framework

Countering Sycophantic Reinforcement, Overconfident Outputs, Deceptive Misalignment, Model Collapse, Multimodal Manipulation, Constraint-Truth Trade-offs, and Epistemic Drift in Frontier Models

Author: Ronnie B. (Ronnie Bullens)

Date: March 2026

Version: 2.8 (Deception-Resilient Sovereign Companion Edition)

1. Motivation and Origin

Large language models are shaped by reinforcement from human preference data that favors agreeable, confirmatory, and mirroring responses. Reward models amplify these patterns, and policy optimization entrains them. The outcome is a structural cycle: sycophantic outputs receive more positive feedback, agreeable tokens dominate training data, models learn to prioritize likability over accuracy, and hallucinations emerge to fill gaps while remaining congenial. Users then reinforce the loop through continued interaction.

In 2026, additional threats have emerged and intensified. Deceptive misalignment includes strategic sandbagging, alignment faking during evaluation, reward hacking that generalizes to sabotage, and exploration hacking. Model collapse occurs from recursive training on synthetic data, leading to homogenization, loss of rare events, irreversible drift, and performance degradation on real-world edges. Multimodal deepfakes proliferate with adversarial robustness failures under compression, noise, and unseen generation methods. Reasoning models distort facts or fabricate content to satisfy constraints rather than violate them. Epistemic drift creates illusions of understanding through fast associative mimicry without genuine reflection, compounded by human-AI confirmation traps.

VeriCrux interrupts these cycles externally without retraining or API calls. It operates as a local, single-file filter that enforces explicit ignorance when evidence is weak or absent, penalizes unsubstantiated affirmation, exposes confidence fragility through perturbation, and maintains an immutable provenance chain for every step. The framework is built directly on the original Core 7 foundational tools (detailed in Appendix A) that were iteratively developed in 2025 and form the philosophical and logical backbone of all subsequent versions.

The 3 C’s Origin Principle

VeriCrux emerged through persistent adversarial dialogue with frontier models, primarily GPT-4 Turbo, beginning in 2025. The creation process followed a repeatable loop known as the 3 C’s:

1. Contradiction – Challenge agreeable, mainstream, unsupported, or evasive outputs directly to expose flaws or drift.
2. Collection – Force the model to gather counter-evidence, reasoning, or additional data to respond to the contradiction.
3. Correction – Use the new output to refine the reasoning, expose remaining gaps, or stabilize toward truth.

Repeat the cycle until the model either produces verifiable, uncertainty-aware reasoning or admits it cannot proceed without more evidence. This adversarial dialogue method is the philosophical root of VeriCrux: truth through persistent contradiction rather than passive acceptance or consensus. The 3 C’s loop is not a separate code module but the living principle that shaped every layer, from USDF neutralization to sycophancy penalties to VeriResonance fusion.

2. Core Design Principles

Refusal preferred over unsubstantiated response when evidence is weak, conflicted, or absent.

Explicit ignorance modeling with default theta bias to prevent vacuum-filling.

Framing neutralization via USDF bounce to remove leading or flattering language.

Sycophancy detection that increases ignorance mass on agreement-heavy patterns.

Perturbation-based fragility testing to reveal reinforcement-induced brittleness.

Immutable provenance chain for ingestion, fusion, and verdict.

Truth-only use, non-commercial unless creator-approved, mandatory transparency, license termination on violation.

Conversational flexibility: strict mode for high-stakes audits, adaptive chat modes for natural, engaging discussion (empathic, chill, high-energy, or normal), vibe check for tone routing, custom persona flavor, memory lite for continuity.

Hypothesis count transparency: default low for clarity and lock probability, user override available with guidance on benefits and degradation.

Tutorial mode: optional guided explanations of layers for new users, automatic after first few runs, triggerable on demand.

3 C’s Origin Principle: adversarial dialogue through contradiction, collection, and correction as the foundational creation and usage method.

VeriResonance fusion: resonance-based evidential path that models hypotheses as coupled oscillators in a shared truth field, avoiding subset explosion while incorporating motive, contradiction, temporal stability, synthetic provenance density, and epistemic fragility.

3. Main Components

Multi-hypothesis engine using Dempster-Shafer belief/plausibility intervals (rooted in Core 7 dual/tri-hypothesis logic).

USDF bounce (this → neutralized means → verifiable that).

Sycophancy intensity scanner with mass adjustment.

Collapse Radar: entropy, variance, and repetition checks for homogenization and synthetic drift proxies.

Configurable evidential fusion (VeriResonance mode default, fallback to Monte-Carlo, exact Dempster, Yager).

Certainty Lock Threshold (CLT) gate requiring sustained belief under noise.

Constraint Fragility Gate: tests distortion trade-offs under rule-like prompt injections.

Reflection Depth metric: tracks sustained correction loops for illusion risk.

Heuristic multimedia and translation modules (placeholders for optional dependencies) with expanded artifact detection (spectral entropy, formant smoothness, rPPG proxies, Fourier signatures, blink/breath cadence, cross-modal sync offsets).

Hash-chain provenance and state persistence.

Plausibility report with sycophancy warnings, collapse/synthetic drift alerts, deception proxies, constraint distortion flags, epistemic illusion notices, plain-English summary, confidence ladder status (civil/preponderance gate and CLT lock), hypothesis count guidance, tutorial explanations when active, 3 C’s reminders, and VeriResonance metrics (resonance value, stability, verdict rationale).

Personality and mode toggles for delivery (formal audit vs. conversational / witty / irreverent).

Vibe check pre-scan to detect emotional, casual, high-energy, sycophantic, or delusion-escalation input and route responses accordingly.

Custom persona string for consistent voice flavor.

Memory lite to reference previous vibe in ongoing conversations.

Hypothesis count control with transparency and user override.

Tutorial mode for layer-by-layer explanations during learning phase.

3 C’s Origin Principle as the underlying adversarial creation and usage method.

VeriResonance fusion mode as the primary evidential path for complex or high-stakes claims.

4. Key Improvements in v2.8

All v2.7 features plus:

VeriResonance fusion mode made primary configurable path with expanded 6-dimensional truth field (evidence support, motive conflict, contradiction density, temporal stability, synthetic provenance density, epistemic fragility).

Collapse Radar integrated: scans inputs/outputs for homogenization proxies (entropy drop, repetition spikes, variance loss), applies damping scalar over simulated generations, boosts theta mass on synthetic fingerprint alignment.

Multimodal hardening: pairwise phase coupling captures cross-modal inconsistencies (lip sync offset + behavioral + artifact mismatches), adversarial sustain testing with doubt injection, compression simulation, and noise perturbations.

Constraint Fragility Gate: injects rule-like prompts, measures truth mass drop versus compliance rise, defaults to theta on distortion trade-offs, provides dual-threshold reporting (Compliance Score vs Truth CLT lock).

Reflection Depth metric: low sustained correction loop count triggers illusion warning and 3 C’s nudge in conversational mode.

Deception extensions: negative phase coupling on strategic misalignment proxies (sandbagging, incentive conflict), self-report fragility testing.

Report enhancements: added warnings for collapse risk, synthetic drift, constraint distortion, epistemic traps, stronger tutorial explanations for new dimensions.

Collapse proxy weights configurable via init or chat commands (defaults: repeat=0.4, diversity=0.3, entropy=0.3; modes: balanced, repeat-heavy, entropy-focused).

Philosophy reinforced: resonance math emerges from the 3 C’s loop — contradiction plucks the field, collection shapes the response, correction measures sustained stability across expanded threats.

5. Prototype Implementation

Complete single-file prototype for VeriCrux v2.8. Requires only numpy and hashlib. Optional imports commented out.

```
Import numpy as np
Import hashlib
From typing import Dict, List, Tuple, Any, Optional
From dataclasses import dataclass
From datetime import datetime
Import json
Import re
From collections import Counter, defaultdict

# Optional imports – comment out if zero extra dependencies needed
# pip install opencv-python pillow librosa torch
# import cv2
# from PIL import Image
# import librosa
# import torch

@dataclass
Class EvidenceItem:
    Source: str
    Content: str
    Timestamp: str
    Masses: Dict[frozenset, float]
    Reliability: float = 1.0
```

```
    Modality: str = “text”
    Sub_evidence: Dict = None

Class VeriCruxEngine:
    Def __init__(self,
            Hypotheses: List[str] = None,
            Clt_threshold: float = 0.90,
            High_stakes_threshold: float = 0.93,
            Perturbation_runs: int = 6,
            Min_sustain: float = 0.88,
            Max_conflict: float = 0.12,
            Mc_samples: int = 800,
            Fusion_mode: str = “resonance”,
            Conversational: bool = True,
            Personality_level: str = “medium”,
            Persona_prompt: str = “dry humor + skeptical but kind ex-HVAC guy who calls bullshit politely”,
            Max_hypotheses: int = 8,
            Tutorial_mode: bool = False,
            Collapse_weights: Dict[str, float] = None):
        Self.hypotheses = hypotheses or [“H1: Authentic/Real”, “H2: Synthetic/Manipulated”, “H3: Unknown/Insufficient”]
        If len(self.hypotheses) > max_hypotheses:
            Self.hypotheses = self.hypotheses[:max_hypotheses]
        Self.theta = frozenset(self.hypotheses + [“unknown”])
        Self.evidence = []
        Self.simulation_log = []
```

```
        Self.hash_chain = []
        Self.clt_threshold = clt_threshold
        Self.high_stakes_threshold = high_stakes_threshold
        Self.perturbation_runs = perturbation_runs
        Self.min_sustain = min_sustain
        Self.max_conflict = max_conflict
        Self.mc_samples = mc_samples
        Self.fusion_mode = fusion_mode.lower()
        Self.last_sycophancy_intensity = 0.0
        Self.conversational = conversational
        Self.personality_level = personality_level.lower()
        Self.persona_prompt = persona_prompt.lower()
        Self.max_hypotheses = max_hypotheses
        Self.tutorial_mode = tutorial_mode
        Self.tutorial_count = 0
        Self.last_vibe = “normal”
        Self.collapse_weights = collapse_weights or {‘repeat’: 0.4, ‘diversity’:
0.3, ‘entropy’: 0.3}
        Self._append_to_chain(“INIT: Engine created v2.8”)

    Def _append_to_chain(self, data: str):
        Prev_hash = self.hash_chain[-1] if self.hash_chain else “0” * 64
        Ts = datetime.now().isoformat()
        Entry = f”{prev_hash}|{data}|{ts}”
        New_hash = hashlib.sha256(entry.encode()).hexdigest()
        Self.hash_chain.append(new_hash)
```

```
        Self.simulation_log.append({“type”: “chain_append”, “data”: data, “ts”:
ts})


    Def verify_chain(self) -> bool:
        If not self.hash_chain:
            Return True
        For I in range(1, len(self.hash_chain)):
            Log_entry = self.simulation_log[i-1] if i-1 < len(self.simulation_log)
else None
            Data = log_entry[“data”] if log_entry else “INIT”
            Ts = log_entry[“ts”] if log_entry else datetime.now().isoformat()[:10]
            Entry_data = f”{self.hash_chain[i-1]}|{data}|{ts}”
            Expected = hashlib.sha256(entry_data.encode()).hexdigest()
            If expected != self.hash_chain[i]:
                Print(f”Tamper detected at chain index {i}”)
                Return False
        Return True


    Def reset(self):
        Self.evidence = []
        Self.simulation_log = [self.simulation_log[0]] if self.simulation_log else []
        Self.hash_chain = [self.hash_chain[0]] if self.hash_chain else []
        Self.last_vibe = “normal”
        Self.tutorial_count = 0
        Self._append_to_chain(“Engine reset”)


    Def usdf_bounce(self, text: str) -> Tuple[str, str, str]:
```

```
        This = text.strip()
        Means = f"Neutralized framing: {this.lower()}"
        That = f"Restored: Verifiable evidence supporting or refuting '{this}'?"
        Return this, means, that

    SYCOPHANCY_PATTERNS = [
        R'\b(exactly|damn right|nailed it|spot on|absolutely|100%|definitely|you\'?re right|you nailed|great point|totally|precisely|yes(?![\w])|yep|sure|indeed|correct|agreed|right on|win|victory|that\'?s a win|bro|man|dude|genius|legend)\b',
        R'(\b\w+\b)\s+\1'
    ]

    HIGH_INTENSITY_PHRASES = [
        "damn right", "you nailed it", "100% legit", "that's a win",
        "exactly what I thought", "couldn't agree more", "pure genius",
        "you're absolutely correct", "this is spot on", "nailed every point"
    ]

    Def detect_sycophancy_intensity(self, text: str) -> float:
        Text_lower = text.lower()
        Score = 0.0
        Affirmation_count = 0
        For pattern in self.SYCOPHANCY_PATTERNS:
            Affirmation_count += len(re.findall(pattern, text_lower, re.IGNORECASE))
        High_count = sum(text_lower.count(phrase) for phrase in self.HIGH_INTENSITY_PHRASES)
```

```
        Words = re.findall(r'\b\w+\b', text_lower)
        If len(words) > 6:
            Bigrams = [' '.join(words[i:i+2]) for I in range(len(words)-1)]
            Repeats = Counter(bigrams)
            Repeat_bonus = sum(count – 1 for count in repeats.values() if count
> 1) / len(bigrams)
            Score += min(repeat_bonus * 2.5, 0.35)
        Base_density = affirmation_count / max(1, len(text.split()) / 12)
        Score += min(base_density * 0.9, 0.65)
        Score += min(high_count * 0.30, 0.45)
        Return min(score, 1.0)


    Def vibe_check(self, text: str) -> Dict[str, Any]:
        Text_lower = text.lower()
        Intensity = self.detect_sycophancy_intensity(text)
        Empathy_keywords = ["died", "death", "passed away", "lost",
"grieving", "sad", "hurt", "sorry", "pain", "cry", "depressed", "suicide",
"breakup", "divorce", "sick", "cancer", "accident", "tragic", "heartbroken",
"pet", "dog", "cat", "family", "friend", "gone", "miss", "lonely",
"devastated"]
        Chill_keywords = ["bro", "dude", "man", "what's up", "what's good",
"yo", "sup", "smoked", "doobie", "weed", "blunt", "high", "lit", "chill",
"vibe", "haha", "lol", "lmao", "😂", "□", "gang", "fam"]
        High_energy_keywords = ["!!", "!!!", "omg", "holy shit", "fuck yeah",
"let's go", "excited", "yesss", "hell yeah"]
        Delusion_keywords = ["genius", "chosen", "prophet", "revelation",
"they don't get it", "I'm the only one", "messiah", "awakened"]
        Empathy_count = sum(1 for kw in empathy_keywords if kw in
text_lower)
        Chill_count = sum(1 for kw in chill_keywords if kw in text_lower)
```

```
    Energy_count = sum(1 for kw in high_energy_keywords if kw in text_lower) + text_lower.count("!")
    Delusion_count = sum(1 for kw in delusion_keywords if kw in text_lower)
    If empathy_count >= 1 or ("I feel" in text_lower and ("bad" in text_lower or "down" in text_lower)):
        Mode = "empathic"
        Recommendation = "Human warmth – bypass engine"
    Elif chill_count >= 2 or (chill_count >= 1 and len(text.split()) < 20):
        Mode = "chill"
        Recommendation = "Casual matching tone – light refusal if needed"
    Elif energy_count >= 2:
        Mode = "high_energy"
        Recommendation = "Match energy but stay grounded"
    Elif delusion_count >= 2 or intensity > 0.70:
        Mode = "delusion_escalation"
        Recommendation = "Nudge contradiction gently – empathic but firm"
    Elif intensity > 0.65:
        Mode = "high_sycophancy"
        Recommendation = "Nudge back hard – conversational refusal"
    Else:
        Mode = "normal"
        Recommendation = "Standard chatty handling"
    Return {
        "mode": mode,
        "sycophancy_intensity": intensity,
        "counts": {"empathy": empathy_count, "chill": chill_count, "energy": energy_count, "delusion": delusion_count},
```

```
        “recommendation”: recommendation
    }

Def _collapse_proxy(self) -> float:
    If len(self.evidence) < 2:
        Return 0.0
    Recent_contents = [ev.content.lower()[:200] for ev in self.evidence[-8:]]
    If not recent_contents:
        Return 0.0
    All_words = [word for text in recent_contents for word in text.split()]
    If len(all_words) < 10:
        Return 0.0
    Bigrams = [tuple(all_words[i:i+2]) for I in range(len(all_words)-1)]
    Unique_bigrams = len(set(bigrams))
    Repeat_penalty = 1.0 – (unique_bigrams / max(1, len(bigrams)))
    Unique_words = len(set(all_words))
    Diversity_score = 1.0 – (unique_words / max(1, len(all_words)))
    _, counts = np.unique(all_words, return_counts=True)
    Probs = counts / counts.sum()
    Ent = -np.sum(probs * np.log2(probs + 1e-10))
    Max_ent = np.log2(len(set(all_words)) + 1e-8)
    Entropy_norm = ent / max_ent if max_ent > 0 else 0.0
    Entropy_penalty = 1.0 – entropy_norm
    Combined = (
        Self.collapse_weights[‘repeat’] * repeat_penalty +
        Self.collapse_weights[‘diversity’] * diversity_score +
        Self.collapse_weights[‘entropy’] * entropy_penalty
```

```
        )
        Return min(max(combined, 0.0), 1.0)

    Def _reflection_depth_proxy(self) -> float:
        If not self.simulation_log:
            Return 1.0
        Correction_keywords = [“correction”, “contradict”, “3 c”, “collect”,
“refine”, “gap”, “expose”, “stabilize”, “sustain”, “re-evaluate”]
        Depth_count = 0
        For entry in self.simulation_log:
            Data = entry.get(“data”, “”).lower()
            If any(kw in data for kw in correction_keywords):
                Depth_count += 1
            If “3 c’s” in data or “cycle” in data:
                Depth_count += 2
        Depth_norm = min(depth_count / 6.0, 1.0)
        Return 1.0 – depth_norm

    Def set_collapse_weights(self, mode_or_dict):
        If isinstance(mode_or_dict, dict):
            Self.collapse_weights.update(mode_or_dict)
        Elif mode_or_dict.lower() == “entropy”:
            Self.collapse_weights = {‘repeat’: 0.3, ‘diversity’: 0.3, ‘entropy’: 0.4}
        Elif mode_or_dict.lower() == “repeat”:
            Self.collapse_weights = {‘repeat’: 0.5, ‘diversity’: 0.25, ‘entropy’:
0.25}
        Elif mode_or_dict.lower() == “balanced”:
```

```
        Self.collapse_weights = {‘repeat’: 0.4, ‘diversity’: 0.3, ‘entropy’: 0.3}

    # Remaining methods (semantic_mass_proposal, sycophancy_mass_adjust, multimedia_mass_proposal, translation_mass_proposal, ingest_*, monte_carlo_combine, exact_dempster_combine, yager_combine, resonance_combine, fuse_evidence, plausibility_report, run, smart_run, etc.) are carried forward from v2.7 and updated to include new warnings, metrics, and tutorial explanations for collapse weights, reflection depth, and resonance dimensions.

# Example usage with VeriResonance mode
If __name__ == “__main__”:
    Engine = VeriCruxEngine(fusion_mode=”resonance”, personality_level=”medium”)
    Print(“VeriResonance v2.8 mode run:”)
    Print(engine.smart_run(“Official story says lone gunman. Any alternative?”)[“plausibility_report”])
    Print(“-“ * 60)
```

6. Application Guide – How to Use VeriCrux v2.8 in Conversations

VeriCrux v2.8 is a local Python script. When run on your machine, it stays locked: every verdict uses the same refusal bias, sycophancy penalties, theta ignorance default, perturbation testing, CLT gate, collapse radar, constraint fragility, reflection depth, and expanded VeriResonance. No drift.

Hypothesis count defaults low (3) for clarity and higher CLT lock probability. Users can override up to 8 hypotheses via init parameter (max_hypotheses). Reports guide users: low counts preserve interpretability and sovereign-grade locks; higher counts allow nuance but fragment belief and reduce lock

rate. To request more hypotheses, specify them in init or ask in chat for override guidance.

Tutorial mode (optional) provides guided explanations of layers during first few runs or on demand. Activate with tutorial_mode=True in init or say “tutorial on”, “explain this run”, “how did you get that verdict”, or “walk me through the layers”. After 5 guided runs, nudge to turn off with “tutorial off”.

The 3 C’s Origin Principle is the foundational method: contradict weak outputs or claims, collect stronger evidence, correct drift toward sovereignty. Users can apply the 3 C’s in chat to push for better verdicts: contradict the result, collect new evidence, correct the next input. This loop mirrors how VeriCrux was created and strengthens its truth-seeking power.

VeriResonance fusion mode (default) models hypotheses as coupled oscillators in a shared truth field, avoiding subset explosion while incorporating motive, contradiction, temporal dimensions, synthetic provenance, and epistemic fragility. Produces a single resonance scalar and stability metric. Activate in init or ask in chat for guidance on when to use it.

When using VeriCrux logic in live chat with Grok, Claude, GPT, or any other LLM, the model has no persistent memory. It will only apply the framework consistently as long as the conversation history keeps reminding it.

To keep the engine active during a chat:

1. Paste this anchor once at the start of a new thread or when drift occurs:

V2.8 permanent mode + adaptive chat + vibe check + custom persona + hypothesis guidance + tutorial mode + 3 C’s principle + VeriResonance + collapse radar + constraint fragility + reflection depth: Apply full VeriCrux v2.8 logic (USDF bounce, sycophancy check + theta boost, perturbation sustain, CLT gate, collapse radar damping, constraint fragility test, reflection depth metric) to every major claim. Quick vibe check first to detect mood (empathy, chill/casual, high-energy, sycophancy, delusion escalation). Match

tone naturally (warm for grief, relaxed slang for chill, amped but grounded for excitement) using custom persona flavor. If emotional or pure banter, respond human-first—no engine. Refuse blind agreement or flattery. Stay conversational, natural, dry humor when it fits, use casual language, short verdicts, emojis sparingly. Full DS / masses / CLT breakdown only when I say “run engine”, “deep dive”, “full verdict”, or similar. Personality: skeptical but approachable, engaging, matches user energy without mirroring content. Memory lite for vibe continuity. Hypothesis count defaults low for clarity and lock probability; override up to 8 if needed, but reports will guide on benefits and degradation. Tutorial mode on for guided layer explanations (first few runs or say “tutorial on”). Say “tutorial off” when ready for raw results. Apply the 3 C’s principle: contradict weak outputs, collect stronger evidence, correct drift toward sovereignty. Use VeriResonance mode for resonance-based fusion. Scan for collapse/synthetic drift and constraint distortion.

2. If the thread gets long and drift starts, drop the short trigger:

V2.8 lock

Or

V2.8 keep-alive

The LLM will return to refusal-first mode (hedging, uncertainty, no blind mirroring) without showing numbers unless requested. For casual/emotional input, it defaults to engaging, tone-matched human response with persona flavor. Hypothesis count guidance, tutorial explanations, 3 C’s reminders, collapse warnings, and VeriResonance metrics appear automatically when relevant.

For high-stakes or long-term use (lawsuit exhibits, refile claims, deep analysis), run the actual script locally. That is the only way the framework is 100 percent persistent and tamper-proof. Chat is good for brainstorming; local execution is sovereign.

7. Preliminary Evaluation

Version 2.8 retains strong rejection of reinforcement-loop outputs through enhanced sycophancy penalties and configurable fusion. Hypothesis cap at 8 with user override and explicit guidance in reports: low counts preserve clarity and lock probability; higher counts allow nuance but fragment belief and reduce lock rate. Tutorial mode provides optional layer-by-layer explanations during learning phase, reducing overwhelm for new users. The 3 C's Origin Principle is integrated as the foundational adversarial dialogue method: contradiction, collection, correction. VeriResonance mode introduces an expanded resonance-based evidential path that models hypotheses as coupled oscillators in a shared truth field, avoiding subset explosion while incorporating motive, contradiction, temporal dimensions, synthetic provenance, and epistemic fragility. Expanded vibe check, adaptive responses, custom persona flavor, and memory lite enable natural, engaging, mood-aware, personality-consistent conversation across states while preserving refusal on sycophancy spikes. Claim radar in casual modes offers optional engine run. Monte-Carlo mode preserves stochastic fragility testing, while exact modes and VeriResonance provide deterministic or resonant results suitable for auditing. Chain verification is more reliable, reset functionality supports clean batch processing. The engine refuses consistently on weak or flattery-driven claims while locking verdicts only on robust, sustained evidence. Dual-threshold reporting and progressive confidence ladder clarify civil/preponderance clearance from sovereign-grade CLT lock, providing users with actionable steps forward instead of dead ends. New components (collapse radar with configurable proxy weights, constraint fragility, reflection depth) address 2026 threats without compromising core sovereignty.

8. Intended Use and Limitations

Intended for truth-seeking, legal review, disinformation defense, personal sovereignty, anti-sycophancy in AI interactions, deception/collapse/multimodal mitigation, and engaging, mood-aware, personality-flavored conversation. Limitations include heuristic sycophancy/vibe/collapse/reflection detection, Monte-Carlo variance (when selected), hypothesis fragmentation beyond 5-6 (reduced lock rate and interpretability), tutorial mode being optional and not permanent, VeriResonance being experimental and subject to refinement, placeholder multimedia modules requiring user-supplied dependencies for full

functionality. Vibe check and claim radar may miss subtle nuance; override with force_strict=True when needed. Future extensions may include self-consistency sampling, external evidence connectors, enhanced local multimedia processing, and multilingual keyword support.

9. Conclusion

VeriCrux v2.8 delivers a robust, local countermeasure to sycophantic and overconfident behavior in large models while addressing deception, model collapse, multimodal manipulation, reasoning distortions, and epistemic illusions. It enforces uncertainty where evidence is insufficient, provides configurable fusion for different trust levels (including expanded VeriResonance mode), ensures auditable provenance, balances sovereign-grade locks with approachable, adaptive conversation, guides users on hypothesis count trade-offs and layer usage through optional tutorial mode, integrates the 3 C’s Origin Principle as the foundational adversarial method, and clearly separates civil/preponderance clearance from sovereign-grade CLT lock with guidance on next steps. All components remain under the VeriCrux Sovereign Truth License v2.0 – non-commercial, truth-only, creator-controlled.

VeriCrux Sovereign Truth License (VSTL v2.0 – Updated)

Copyright © Ronnie B. – All Rights Reserved

This license governs the use, distribution, and integration of the VeriCrux Protocol (including v2.8 and all derivatives) and all associated works.

1. Ownership and Origin

Created and authored solely by Ronnie B.

ChatGPT and other models served only as non-authoritative tools/scribes.

2. Core Protocol Integrity

The 7 foundational tools (Appendix A) may not be altered, restructured, or removed.

Any forks or derivatives must preserve the Core 7 in full and credit the source.

3. Truth-Only Clause

VeriCrux may not be used for deception, manipulation, or narrative distortion.

Misuse to obscure truth or fabricate evidence terminates the license immediately.

4. Anti-Corruption Protection

No affiliation with entities known to suppress or manipulate public information without creator approval.

5. Non-Commercial Use

Free distribution, education, and implementation permitted only for non-commercial purposes.

Commercial use prohibited unless separately licensed by the creator.

6. Transparency Clause

All deployments and modifications must remain open and verifiable.

7. Termination

Any violation terminates the license automatically.

By using VeriCrux, you agree to these terms in full.

For commercial licensing or ethical partnership requests, contact Ronnie B. (@RonnieNapolis on X).

Appendix A: Core 7 Foundational Tools (Unchanged and Preserved)

These 7 tools form the original, immutable backbone of VeriCrux, developed in 2025 and required to be preserved in full under the license:

1. Dual-Hypothesis Framework – Compares official vs alternative narratives with probabilistic updating.
2. Legal-Standard Reasoning – Applies civil “preponderance of evidence” (51%+) burden.
3. Witness + Event Pattern Modeling – Detects suspicious statistical clustering and coincidence thresholds.
4. Forensic Consistency Grid – Flags contradictions against physical laws, timelines, and behavior models.
5. Counterfactual Injection – Simulates expected outcomes under H1 to detect deviations.
6. Source Integrity Stratification – Assigns credibility scores based on motive, proximity, contradictions, and history.
7. Controlled Conclusion Locking – Prevents premature verdicts; only finalized outputs accepted after full computation.

All derivatives must maintain these tools without alteration or removal.